

**Section III:**  
**AMENDMENT UNDER 37 CFR §1.121 to the**  
**DRAWINGS**

No amendments or changes to the Drawings are proposed.

**Section IV:**  
**AMENDMENT UNDER 37 CFR §1.121**  
**REMARKS**

**Rejections under 35 U.S.C. §102(e)**

In the Office Action, Claims 1 - 5, 10 - 14 and 15 - 18 were rejected as follows:

*Examiner in the Office Action:*

"Claims 1-5, 10 - 14 and 15 - 18' are rejected under 35 U.S.C. 102(e) as being clearly anticipated by Burdick et al (USPG Pub No. 2004/0107203A1)"

Claim 1. The Examiner stated the rationale for rejecting Claim 1 as:

*Examiner in the Office Action:*

"As for Claim 1, Burdick et al teaches "generating a set of cleaning attributes for each cleaned data record in a complete set of cleaned data records, said cleaning attributes reflecting which fields of each record have been modified by a cleaning operation" (see paragraph [0045] and [0057-0058]); "receiving a data feature identified by a data mining process for a subset of said complete set of cleaned data records" (see paragraph [0038], [0051]); "determining a degree of correlation of said data feature to the modified fields of said subset of cleaned data records according to said cleaning attributes" (see paragraph [0032-00361]; "and declaring said data feature as suspect responsive to said degree if correlation exceeding a threshold" (see paragraph [0035], [0053])."

With respect to the first step of Claim 1 "generating a set of cleaning attributes for each cleaned data record in a complete set of cleaned data records, said cleaning attributes reflecting which fields of each record have been modified by a cleaning operation", Applicant respectfully disagrees that Burdick teaches "cleaning attributes".

By "cleaning attributes", Applicant means (emphasis added by Applicant):

Applicant's Disclosure:

[0026] During the data cleaning process, each "row" or record in the cleaned data set will have been assigned to a cluster. The **cleaning attribute** associated with each cleaned record **indicates which fields in the record have been modified, and which are in original state, preferably in a bit-mapped or "bit flag" register format.**

[0027] At least four embodiments of our "data cleaning flags" are available within the scope of the present invention, including but not limited to:

- (a) maintaining the data cleaning flags as a part of the cleaned data records;
- (b) maintaining the data cleaning flags in a parallel table containing only references to cleaned data records;
- (c) maintaining a parallel table of data cleaning flags which includes a data record key, a cleaned field ID, and possibly the "raw" or pre-cleaned data value;
- (d) maintaining a cleaned field list (f1=y, f5=y, f7=y) in any of the formats described in (a), (b), or (c).

Burdick teaches cleaning of data using rules to govern *how* the cleansing process is performed, but is silent regarding generating any flags or attributes to track which record fields have been modified. The relied-upon portions of Burdick are silent as to such flags or attributes:

Burdick Cited Art Passage:

[0045] An automated learning component 203 **performs the cleansing process**. This component 203 also supports a learning system for refining the cleansing process (i.e., adjusting the algorithm parameters to allow for either a more efficient and/or more accurate solution).

...

[0057] The **rules layer 403** defines the execution of the cleansing process. Each processing layer section has a

corresponding rules layer section in the rules layer 403. Each rules layer section contains the **rules for controlling the execution of the corresponding processing layer section**. For each step of any cleansing process, the rules define the requirements for each step for automated evaluation. For example, the rules controlling the clustering section determine **how the clustering module should build the clusters** for each real-world entity represented in the record collection. [0058] The **rules for each step are derived initially from an execution plan** (given as input to the automated learning component 203), and are refined by input from a learning layer 404 for that step during the data cleansing process. Since each step of the data cleansing process has different requirements, the rules to perform each of the steps may take different forms. **Rules may be given as Boolean expressions, IF-THEN statements, threshold values, etc.**

Burdick's "rules" of how to perform the cleansing process are not the same as creating flags or attributes which track whether or not a field has been modified as a result of a cleansing process. Rather, Burdick's process would simply replace the unmodified field value with the "cleansed" value, and would not create a tracking flag or "cleaning attribute" as defined and described the Applicant.

The term "cleaning attribute" must be given its meaning in the claims according to the Applicant's disclosure, because the claims are part of the disclosure:

**35 U.S.C. 112:**

...

The specification shall conclude with one or more claims particularly pointing out and distinctly claiming the subject matter which the applicant regards as his invention.

...

**Federal Circuit regarding Interpretation of Claim Terms in view of Inventor's Disclosure:**

"Importantly, the person of ordinary skill in the art is deemed to read the

claim term not only in the context of the particular claim in which the disputed term appears, but in the context of the entire patent, including the specification."

...

"For that reason, claims must be read in view of the specification, of which they are part . . . [T]he specification is always highly relevant to the claim construction analysis. Usually, it is dispositive; it is the single best guide to the meaning of a disputed term . . . "

"Consistent with that general principle, our cases recognize that the specification may reveal a special definition given to a claim term by the patentee that differs from the meaning it would otherwise possess. In such cases, the inventor's lexicography governs. . . . In other cases, the specification may reveal an intentional disclaimer, or disavowal, of claim scope by the inventor. In that instance as well, the inventor has dictated the correct claim scope, and the inventor's intention, as express in the specification, is regarded as dispositive." *Phillips v. AWH Corp.*, 415 F.3d 1303, 75 USPQ2d 1321 (Fed. Cir. 2005) (en banc).

For these reasons, Applicant respectfully disagrees that Burdick anticipates this step of Claim 1, and allowance of Claim 1 is requested.

With respect to the second step of Claim 1 "receiving a data feature identified by a data mining process for a subset of said complete set of cleaned data records" being anticipated by Burdick at paragraphs [0038] and [0051], Applicant respectfully disagrees.

By "data feature identified by a data mining process for a subset of cleaned data records", Applicant means an aspect of the data records which has been identified by data mining processes such as a cluster, trend or pattern:

Dictionary:

feature *n.*

...

2. A prominent or distinctive aspect, quality, or characteristic: a feature of one's personality; a feature of the landscape. (*The American Heritage® Dictionary of the English Language*, Fourth Edition. Retrieved August 30, 2007, from Dictionary.com website: <http://dictionary.reference.com/browse/feature>)

Applicant's Disclosure:

[0026] During the data cleaning process, each "row" or record in the cleaned data set will have been assigned to a cluster. The cleaning attribute associated with each cleaned record indicates which fields in the record have been modified, and which are in original state, preferably in a bit-mapped or "bit flag" register format.

...

[0029] A subsequent data mining clustering process is employed to find clusters, and to provides a list of attributes that most influenced individuals becoming members of the cluster. The attribute list is preferably in "entropy" order, meaning that customers in the cluster have a high percentage of this same value, whereas customers outside the cluster have a low percentage of this attribute. Well-known entropy ordering methods use a mathematical ratio such as percentage in a cluster to percentage outside of a cluster (e.g. [% in cluster] / [% outside of cluster] ).

The term "data feature" as widely used in the art of data mining, and in the context of Applicant's disclosure regarding data cleaning and data mining, conveys a generic class of aspects about data which can be discovered and analyzed in a group of data records.

Burdick, however, at paragraphs [0038] and [0051 is only disclosing that their records contain information about a real world entity, in which fields contain data (well known structure of a record), type of data (string, number, data, etc.), record count, source of the record (keyboard, etc.), but is not describing receiving a "data feature" (e.g. cluster, etc.) identified in a subset of data records which have been associated with Applicant's cleaning attributes or flags (emphasis added by Applicant):

Burdick Cited Art Passage:

[0038] **Each record contains information** about a real world entity. Each record can be **divided into fields**, each field describing an **attribute of the entity**. The format of each record includes information about the number of fields in the record and the order of the fields. The format also defines the type of data in each field (for example, whether the field

contains a **string, a number, date, etc.**).

...

[0051] An information generating module 303 of the preprocessing component 202 generates information about the record collection for input into the automated learning component 203. This generated information may be derived from the record collection itself. For example, statistics A , about the record collection (i.e., **how many records come from a particular source, how many records share a particular value for a field**, etc.) may be computed. Indices of different record fields may be built, or the records examined to determine the type of data in each record field (i.e., **whether the data is alphabetic, numeric, a calendar date**, etc.) Available information outside of the record collection itself may also be used. Examples may include how record data was entered (i.e., **whether the record data was taken over the phone, typed into the system at a keyboard, OCR'd into the system, etc.**), the source of the record, or metadata about the record fields.

As such, Applicant respectfully submits that Burdick fails to anticipate the second step of Claim 1, and respectfully requests allowance of Claim 1.

Regarding the third step of Claim 1 "determining a degree of correlation of said data feature to the modified fields of said subset of cleaned data records according to said cleaning attributes", it was reasoned in the Office Action that Burdick anticipates this step at paragraphs 0038 and 0051. Applicant respectfully disagrees.

First, in order to determine a correlation between a set of modified fields and a data feature, a process must somehow have knowledge of which fields were changed during cleaning. Since Burdick does not disclose flags or attributes to track which fields are modified during cleaning, Burdick is unable to, and silent regarding, determining a correlation between changed/cleaned fields and a data feature identified by a data mining process.

Referring to the specific paragraphs relied upon in the reasoning 0038 and 0051, which are quoted in the foregoing paragraphs, Burdick merely states that their records contain information, their records are divided into fields, and their preprocessing component may

perform some statistical (not mining) processes on these records, such as counting the records, counting how many records share a value, etc.

For these reasons, Applicant respectfully submits that Burdick fails to anticipate the third step of Claim 1, and allowance of Claim 1 is respectfully requested.

With respect to the fourth step of Claim 1 ""and declaring said data feature as suspect responsive to said degree of correlation exceeding a threshold" being anticipated by Burdick's paragraphs 0035 and 0053, Applicant respectfully disagrees. Whereas Burdick is unable to determine a degree of correlation between a data feature found by data mining due to not having any flags or attributes which indicate which fields were modified by data cleaning methods, Burdick is further unable to declare that a data feature is suspect (e.g. potentially unreliable or false) due to possible false data features occurring as a result of data cleaning changes to the data.

In fact, Burdick's paragraph 0035 does not mention declaring any result as "suspect", but merely describes how there process removes duplicate or "garbage" records (e.g. a form of cleaning, not review of data mining results):

*Burdick's Cited Art Passage:*

[0035] In the clustering and matching steps, algorithms **identify and remove duplicate or "garbage" records** from the collection of records. Determining if two records are duplicates involves performing a similarity test that quantifies the similarity (i.e., a calculation of a similarity score) of two records. If the similarity score is greater than a certain threshold value, the records are considered duplicates.

And, Burdick's paragraph 0053 does not mention declaring any result as "suspect" based on correlation of a data feature to a set of modified and cleaned fields, but instead describes determining whether output from their single-source module, information generating module, and planning module is "satisfactory" (emphasis added by Applicant):



Burdick's Cited Art Passage:

[0053] An output evaluation module 305 of the pre-processing component 202 evaluates the output of the three other functional modules 302, 303, 304. If the output is **determined to be satisfactory**, the output is passed to the automated learning component 203 via an output module 306. **If the output is determined to be unsatisfactory (i.e., based on pre-defined thresholds, application-specific metrics, etc.)**, the three other functional modules 302, 303, 304 may be run again with different parameters. The output evaluation module 305 also may provide suggestions on how to change the execution of the three other functional modules 302, 303, 304 to improve the quality of the output (i.e., a feedback loop).

Burdick is silent regarding how the output is determined to be "satisfactory" or "unsatisfactory". A word search of the entire disclosure reveals no other instances of these two terms. Further, a word search of the term "threshold" only reveals two more instances, one regarding how to determine if records are duplicates of each other in para. 0035, the other regarding the "rules" for executing their cleaning algorithms (para. 0058), both of which precede mining processes and thus could not be reviewing or declaring anything about the mining output.

For these reasons, Applicant respectfully submits that Burdick fails to anticipate this step of Claim 1, and allowance of Claim 1 is requested.

Claim 2. The rationale for rejecting independent Claim 2 was set forth in the Office Action as follows:

Examiner in the Office Action:

As for Claim 2, Burdick et al teaches "generating a set of bit-mapped Boolean flags to form a cleaning attributes register for each cleaned data record" (see paragraph [0057-0058]).

Applicant respectfully disagrees. Burdick is referring to "Boolean expressions", not flags, as part of their cleaning rules, not as part of their data cleaning structures (emphasis added by Applicant):

Burdick's Cited Art Passage:

[0058] The rules for each step are derived initially from an execution plan (given as input to the automated learning component 203), and are refined by input from a learning layer 404 for that step during the data cleansing process. Since each step of the data cleansing process has different requirements, the rules to perform each of the steps may take different forms. Rules may be given as **Boolean expressions**, IF-THEN statements, threshold values, etc.

A "Boolean expression" is an evaluation expression, such as testing a value to see if it is true or false, and depending on the results of the test, proceeding to one or more steps:

Dictionary:

Boolean expression

An expression that results in a value of either TRUE or FALSE. For example, the expression

$2 < 5$  (2 is less than 5)

is a Boolean expression because the result is TRUE. All expressions that contain *relational operators*, such as the *less than sign* (<), are Boolean.

Boolean expressions are also called *comparison expressions*, *conditional expressions*, and *relational expressions*. (Source: *Random House Webster's Computer and Internet Dictionary*, Third Edition, by Philip E. Margolis, pg. 61)

A "Boolean flag", however, is not an expression at all, but instead a data value itself, such as "1" or "0":

Dictionary:

**flag** *n.*

1. A software or hardware mark that signals a particular condition or status. A flag is like a switch that can be either on or off. The flag is said to be *set* when it is turned on.
2. A special mark indicating that a piece of data is unusual. For example, a record might contain an *error flag* to indicate that the record consists of unusual, probably incorrect, data.

(Source: *Random House Webster's Computer and Internet Dictionary*, Third Edition, by Philip E. Margolis, pg. 216)

As such, Burdick teaches of Boolean *expressions* in their "rules" for performing data cleaning, but is silent regarding associating Boolean *flags* with the fields which have been modified during cleaning.

For these reasons, and for the reasons discussed regarding the rejection of Claim 1, Applicant requests allowance of Claim 2.

Claim 3. Regarding the claimed step of appending or prepending a set of cleaning attributes (e.g. flags) to each cleaned data record, and generating an attribute table, the Examiner has reasoned that Burdick's disclosure anticipates this claim step at paragraphs [0043-0044] and [0055-0059].

Applicant respectfully disagrees. Burdick's paragraph 0043 pertains to their five components of their architecture, including an input component, but is silent regarding appending or prepending anything to the data records. Burdick's paragraph 0044 pertains to their pre-process component which "prepares" the data to be cleaned, but discloses nothing about prepending or appending operations. Burdick's paragraph 0055 discusses their automated learning component which does the data cleaning and how it receives outputs from three other components, but there is no mention of appending or prepending anything to the records. Likewise, Burdick's paragraphs 0056 - 0059 pertain to their processing layer, rules layer, rules for each step, and the learning layer, respectively, but all are silent regarding any appending or

prepending operations. In fact, a word search of the entire Burdick disclosure reveals no occurrences of the terms "append" or "prepend", nor any possible contextual synonyms such "concatenate" (the term "append" appears once related to reference to the "appended claims" following the disclosure).

For these reasons, Applicant requests allowance of Claim 3 whereas Burdick fails to anticipate the steps and limitations of Claim 1, and further fails to teach the steps and limitations of Claim 3.

Claim 4. It was reasoned in the Office Action that Claim 4 is anticipated by Burdick:

*Examiner in the Office Action:*

Burdick et al teaches "a step selected from the group of receiving a cluster, receiving a trend, and receiving a pattern" (see paragraph [0032-0034], [0065-0067], and [0071]).

Applicant respectfully disagrees. Claim 4 depends from Claim 1, and thus incorporates the steps and limitations untaught by Burdick as discussed in the foregoing paragraphs.

For this reason, Applicant requests allowance of Claim 4.

Claim 5. It was reasoned in the Office Action that Claim 5 is anticipated by Burdick:

*Examiner in the Office Action:*

As for Claim 5, Burdick et al teaches "comparing each record in a raw data set to each record in a cleaned data set" (see paragraph [0069-0070]).

Applicant respectfully disagrees. Claim 4 depends from Claim 1, and thus incorporates the steps and limitations untaught by Burdick as discussed in the foregoing paragraphs.

Burdick's paragraphs 0069 and 0070 disclose that their results evaluator module only receives the cleansed record collection and "additional information from the automated learning component". While they somehow determine if the cleansed data records meet "quality metrics", it is not stated *how* this is determined, and especially is not stated that it is done by comparing the raw (original) records to the modified (cleansed) records.

For these reasons, Applicant respectfully requests allowance of Claim 5.

Claims 10 - 14. It was reasoned in the Office Action that Claims 10 - 14 are anticipated by Burdick:

*Examiner in the Office Action:*

Claims 10-14 differ from Claims 1-5 in that claims 10-14 are computer readable medium whereas claims 1-5 are method claims. Thus, claims 10-14 are analyzed as previously discussed with respect to claims 1-5 above.

Applicant agrees that Claims 10 - 14 are directed towards computer-readable medium embodiments of the invention in an analogous manner to Claims 1 - 5. Applicant respectfully disagrees that Claims 1 - 5 and 10 - 14 are anticipated by Burdick for the reasons set forth in the foregoing paragraphs.

Claims 15 - 18. It was reasoned in the Office Action that Claims 15 - 18 are anticipated by Burdick:

*Examiner in the Office Action:*

Claims 15-1 8 differ from Claims 1-4 in that claims 15-1 8 are system whereas claims 1-4 are method claims. Thus, claims 15-1 8 are analyzed as previously discussed with respect to claims 1-4 above.

Applicant agrees that Claims 15 - 18 are directed towards system embodiments of the invention in an analogous manner to Claims 1 - 4. Applicant respectfully disagrees that Claims 1 - 4 and 15 - 18 are anticipated by Burdick for the reasons set forth in the foregoing paragraphs.

For the reasons set forth herein, Applicant respectfully submits that the cited art fails to clearly anticipate Claims 1 - 4 and 10 - 18 as required under 35 U.S.C. §102(e). Title 35 U.S.C. §102 states "A person shall be entitled to a patent unless . . . ". It is respectfully submitted that Applicant is entitled to a patent on Claims 1 - 4 and 10 - 18.

Respectfully,

*/ Robert Frantz /*

Robert H. Frantz, Reg. No. 42,553  
Agent for Applicant  
Tel: (405) 812-5613  
Franklin Gray Patents, LLC